

Results from the *read-15* test

Adam Lyon, August 1, 2005

1 Introduction

A SAM test was performed to determine the rate of file deliveries possible on a test CAF system.

Appendix A R Session

R is an open source statistical analysis software package that allows for very easy analysis of data in databases and text files. I wrote a "notebook" style package that allows one to use R from within Microsoft Word. Below is the notebook providing all of the code and results for this document.

A.1 Connect to R and initialize

Connect to R running locally on my laptop.

```
<R0> #connect port 6101 timeout 20
R is using work directory /Users/adam/work/projects/cdfSamTests/read-15
Set up graphics
<R1> library(lattice)

<R2> trellis.par.set(col.whitebg())

<R3> fontsize = trellis.par.get("fontsize"); fontsize$text=16 ;
      fontsize$points=6 ; trellis.par.set("fontsize", fontsize)
```

I have a helper function that makes putting graphics into Word easy.

```
<R4> mp
function (plotExpr, file, height = 7, width = 7, res = 72 * 3)
{
  bitmap(file, "pngalpha", height = height, width = width,
         res = res, pointsize = 10)
  r = eval(plotExpr)
  if (class(r) == "trellis")
    print(r)
  invisible(dev.off())
}
<environment: namespace:RemoteRSOAP>
```

A.2 Running job output

Doug's python script loops over getting the next file, waiting thirty seconds to simulate processing, and then releasing the file. A log file records the time of the get next file, the duration of the command, the pnfs name of the file, and the time and duration of the release file command.

Let's read this information into R.

```

<R5> d = read.table("out.log", header=T)

<R6> d[1:5,]
   job segment   getDate   getTime getDur fileNum
1 1305        1 2005-08-01 01:05:51  3.159      1
2 1305        1 2005-08-01 01:06:09  2.069      2
3 1305        1 2005-08-01 01:06:26  2.074      3
4 1305        1 2005-08-01 01:06:43  8.070      4
5 1305        1 2005-08-01 01:07:07 10.095      5

pnfs
1
dcap://cdfdca1.fnal.gov:25140/pnfs/fnal.gov/usr/cdfen/filesets/GJ/GJ22/
GJ2270/GJ2270.0/xd025a3f.0070bhd0
2
dcap://cdfdca1.fnal.gov:25140/pnfs/fnal.gov/usr/cdfen/filesets/GJ/GJ22/
GJ2275/GJ2275.0/xd025b96.014bbhd0
3
dcap://cdfdca1.fnal.gov:25140/pnfs/fnal.gov/usr/cdfen/filesets/GJ/GJ22/
GJ2270/GJ2270.0/xd025982.029bbhd0
4
dcap://cdfdca1.fnal.gov:25140/pnfs/fnal.gov/usr/cdfen/filesets/GJ/GJ22/
GJ2270/GJ2270.0/xd025b54.00d7bhd0
5
dcap://cdfdca1.fnal.gov:25140/pnfs/fnal.gov/usr/cdfen/filesets/GJ/GJ22/
GJ2270/GJ2270.0/xd025a60.0086bhd0
    relDate  relTime relDur
1 2005-08-01 01:06:09  0.219
2 2005-08-01 01:06:26  0.109
3 2005-08-01 01:06:43  0.098
4 2005-08-01 01:07:06  0.078
5 2005-08-01 01:07:32  0.088

```

Convert dates and times into POSIX times that R can deal with...

```
<R7> d$getDateTime = paste(d$getDate, d$getTime)
```

```
<R8> d$relDateTime = paste(d$relDate, d$relTime)
```

```
<R9> d$getPTime = as.POSIXct( strptime( d$getDateTime, "%Y-%m-%d
%H:%M:%S" ) )
```

```
<R10> d$relPTime = as.POSIXct( strptime( d$relDateTime, "%Y-%m-%d
%H:%M:%S" ) )
```

We can make the delivery time from the get time plus the duration

```
<R11> d$delPTime = d$getPTime + d$getDur
```

What is the range of the test?

```
<R12> testEdges = c(min(d$getPTime), max(d$relPTime, na.rm=T)) ;  
      testEdges  
[1] "2005-08-01 01:05:50 CDT" "2005-08-01 10:03:28 CDT"  
  
<R77> testTime = diff(testEdges) ; testTime  
Time difference of 8.960556 hours  
  
<R14> prettyEdges = c(testEdges[1]-60*60, testEdges[2]+60*60)
```

A.2.1 Basics

How many files deliveries were attempted?

```
<R15> nrow(d)  
[1] 39308
```

How many failed on the get end?

```
<R16> sum(is.na(d$getDur))  
[1] 0
```

How many failed on the release end?

```
<R17> sum(is.na(d$relDur))  
[1] 0
```

Merge job and segment numbers

```
<R18> d$jobseg = paste(d$job, d$segment)
```

What were the jobs and segments?

```
<R19> d$jobseg[ is.na(d$relDur) ]  
character(0)  
  
<R20> #var successfulDeliveries = nrow(d)  
39308
```

How come this isn't 40,000?

Look up the maximum file number for each job and segment.

```
<R79> largestFNum = tapply(d$fileNum, d$jobseg, max)
```

How many did not get all 40 files?

```
<R82> length( largestFNum[ largestFNum < 40 ] )  
[1] 143
```

Hmmm.

```
<R83> notDone = largestFNum[ largestFNum < 40 ]  
<R84> notDone[1:3]  
1305 21 1305 31 1305 32  
      38      37      38
```

Will have to look these up!

A.2.2 File delivery rate

```
<R88> fileRatePerDay = nrow(d)/(as.numeric(testTime))*24;  
      fileRatePerDay  
[1] 105282.8
```

In a perfect world, what should be the mean wait time?

```
<R89> filesPerSeg = nrow(d) / length(unique(d$jobseg)) ; filesPerSeg  
[1] 39.308
```

Assume that every segment runs the entire length of the test. Therefore (in minutes),

```
<R90> meanWaitTime = as.numeric(testTime)*60 / filesPerSeg ;  
      meanWaitTime  
[1] 13.67745
```

How many seconds per file?

```
<R95> secondsPerFile = as.numeric(testTime)*60*60 / nrow(d) ;  
      secondsPerFile  
[1] 0.8206472
```

A.2.3 Number of segments running

Let's plot how many segments were running at a given time.
A segment turns on when it does its first "get next file". It turns off when it does its last release.

How many segments are there?

```
<R21> length( unique( d$jobseg ) )  
[1] 1000
```

So we have a record of all 1000 segments. Good!

Segment starts when the first file is requested

```
<R22> segStart = data.frame( time=d$getPTime[d$fileNum==1], adj=1 )
```

```

<R23> segEnd = data.frame( time=d$relPTime[d$fileNum==40], adj=-1 )

<R24> segs = rbind(segStart, segEnd)

<R25> segs = segs[ order(segs$time), ]

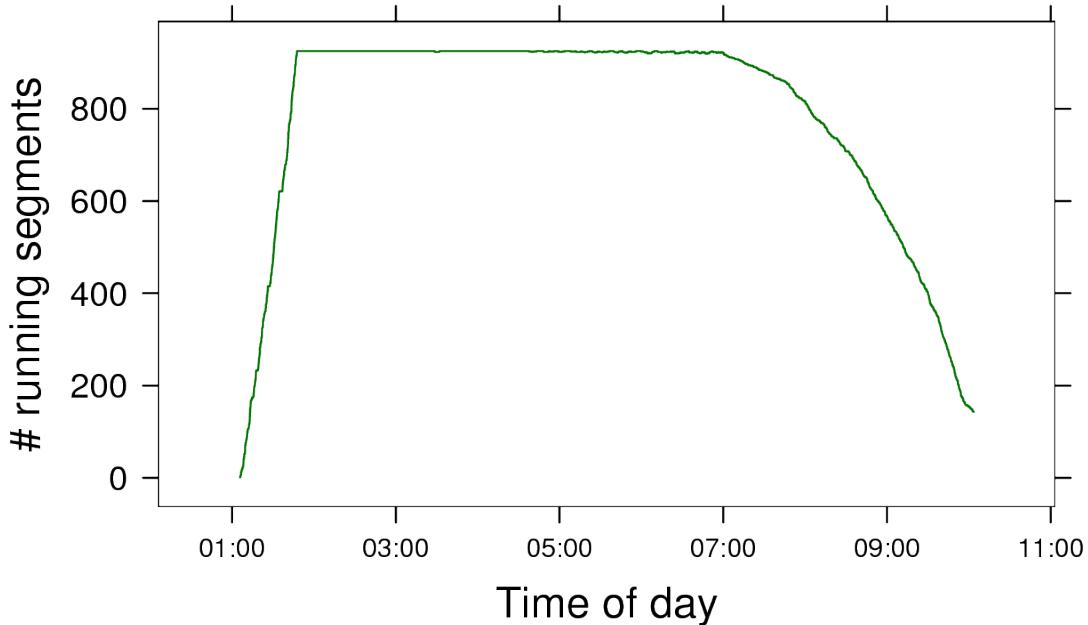
<R26> segs$count = cumsum(segs$adj)

<R27> segs[1:10,]
      time adj count
3 2005-08-01 01:05:50   1    1
1 2005-08-01 01:05:51   1    2
5 2005-08-01 01:06:04   1    3
6 2005-08-01 01:06:07   1    4
223 2005-08-01 01:06:10  1    5
4 2005-08-01 01:06:16  1    6
221 2005-08-01 01:06:18  1    7
7 2005-08-01 01:06:29  1    8
2 2005-08-01 01:06:35  1    9
225 2005-08-01 01:06:36  1   10

<R29> mp(
  xyplot( segs$count ~ segs$time, type="s",
          main="Number of running segments",
          xlab="Time of day",
          ylab="# running segments",
          scales=list(x=list(tick.number=6, cex=0.6)),
          xlim=prettyEdges ),
  "nsegs.png", h=4, w=6 )
#with graphics nsegs.png timeout 120

```

Number of running segments



A.2.4 Number of gets and deliveries

Let's count up how many get file requests and deliveries there were per hour

```
<R30> getsPerHourCuts = cut( d$getPTime, "hours" )

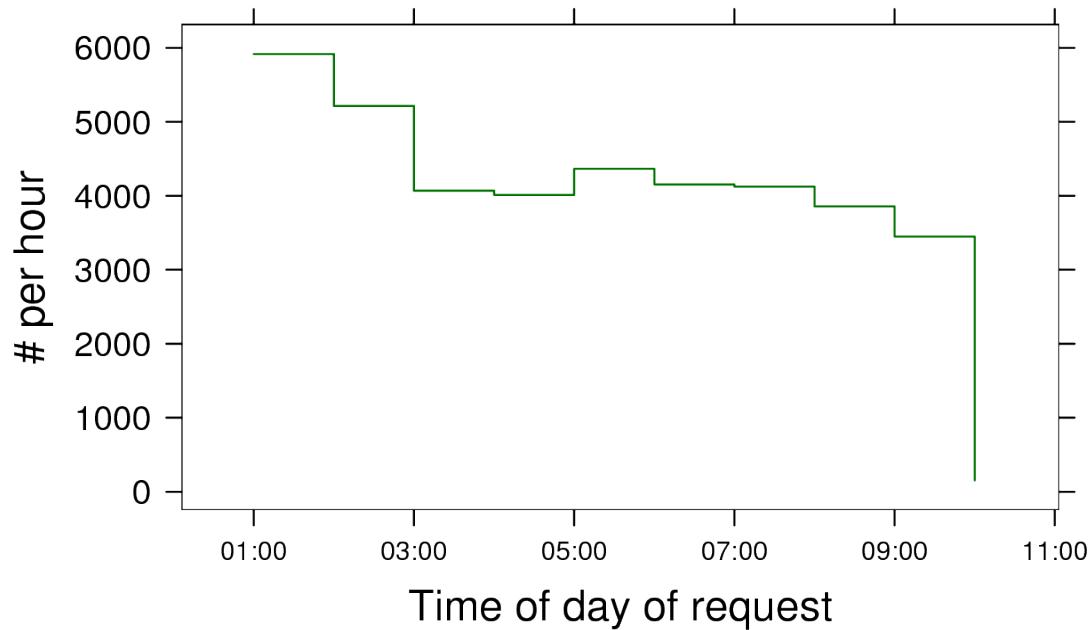
<R31> delsPerHourCuts = cut( d$delPTime, "hours" )

<R32> getsPerHour = tabulate(getsPerHourCuts)

<R33> delsPerHour = tabulate(delsPerHourCuts)

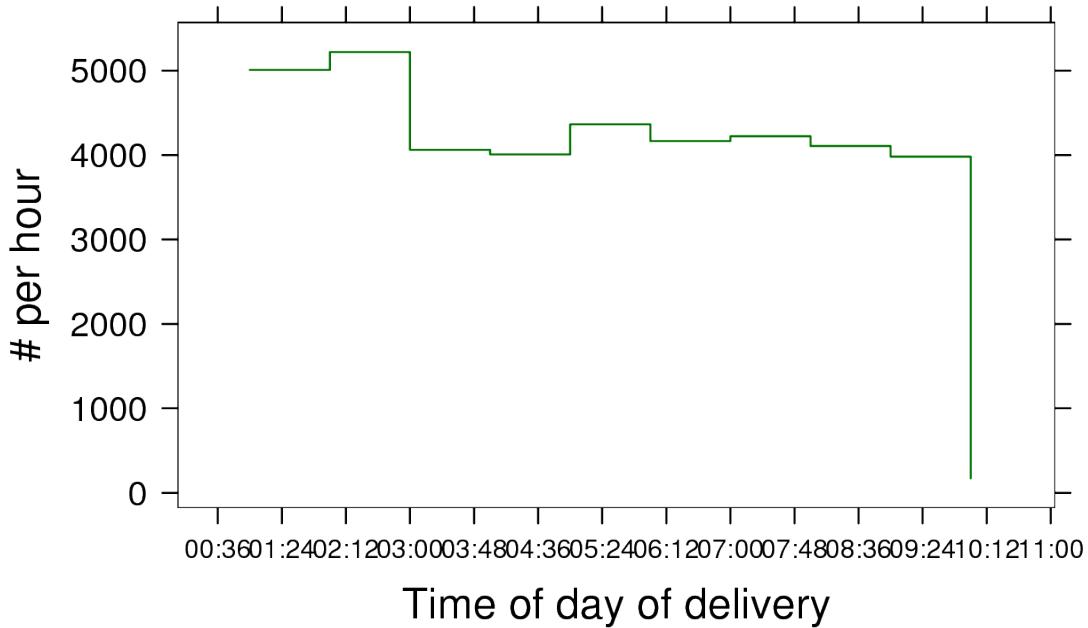
<R35> mp(
  xyplot( getsPerHour ~ as.POSIXct(levels(getsPerHourCuts)),
    main="Get file requests",
    xlab="Time of day of request",
    ylab="# per hour", type="s", xlim=prettyEdges,
    scales=list(x=list(tick.number=6, cex=0.6)))
),
"getsPerHour.png", h=4, w=6
)
#with graphics getsPerHour.png timeout 60
```

Get file requests



```
<R36> mp(
  xyplot( delsPerHour ~ as.POSIXct(levels(delsPerHourCuts)),
    main="File deliveries",
    xlab="Time of day of delivery", xlim=prettyEdges,
    ylab="# per hour", type="s",
    scales=list(x=list(tick.number=6, cex=0.6))
  ),
  "delsPerHour.png", h=4, w=6
)
#with graphics delsPerHour.png timeout 60
```

File deliveries



Let's just get a cumulative plot of when deliveries occurred.

```
<R37> deliveries = data.frame( time=d$delPTime, adj=1 )
```

```
<R38> deliveries = deliveries[ order(deliveries$time), ]
```

```
<R39> deliveries$count = cumsum(deliveries$adj)
```

```
<R40> nrow(deliveries)
```

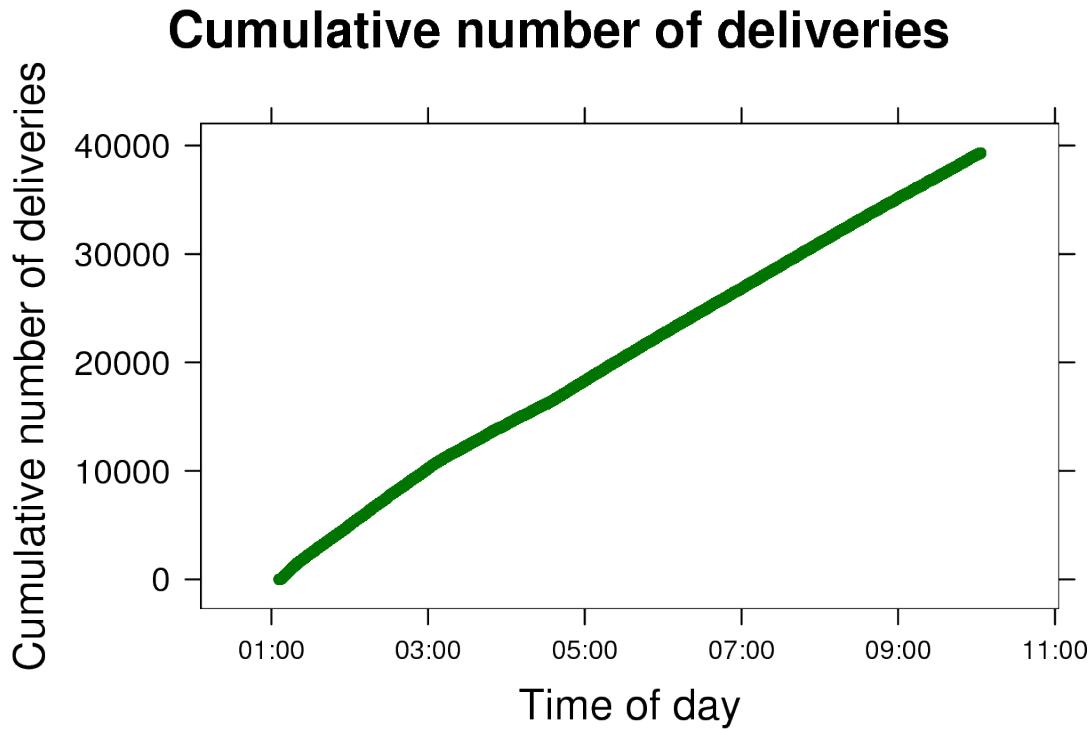
```
[1] 39308
```

```
<R41> deliveries[39270:39281,]
      time adj count
27838 2005-08-01 10:02:08 1 39270
34598 2005-08-01 10:02:10 1 39271
36080 2005-08-01 10:02:10 1 39272
33229 2005-08-01 10:02:12 1 39273
32410 2005-08-01 10:02:12 1 39274
30985 2005-08-01 10:02:12 1 39275
28620 2005-08-01 10:02:14 1 39276
29400 2005-08-01 10:02:15 1 39277
33923 2005-08-01 10:02:16 1 39278
30221 2005-08-01 10:02:17 1 39279
32486 2005-08-01 10:02:18 1 39280
34719 2005-08-01 10:02:18 1 39281
```

```

<R43> mp(
  xyplot( deliveries$count ~ deliveries$time, type="p",
         main="Cumulative number of deliveries",
         xlab="Time of day",
         ylab="Cumulative number of deliveries",
         scales=list(x=list(tick.number=6, cex=0.6)),
         xlim=prettyEdges ),
  "ndel.png", h=4, w=6 )
#with graphics ndel.png timeout 120

```



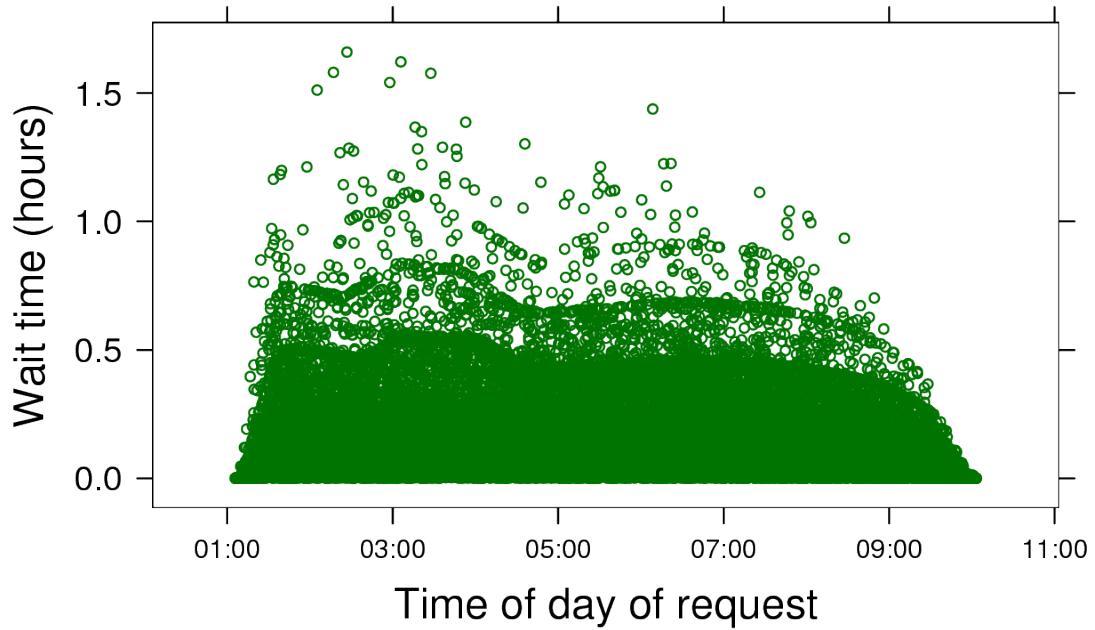
What do the wait times look like?

```

<R45> mp(
  xyplot( getDur/60/60 ~ getPTime, data=d,
         main="File delivery waits",
         xlab="Time of day of request", xlim=prettyEdges,
         ylab="Wait time (hours)",
         scales=list(x=list(tick.number=6, cex=0.6)))
  ),
  "getWaits.png", h=4, w=6
)
#with graphics getWaits.png timeout 60

```

File delivery waits



```
<R46> mp(
  xyplot( getDur/60/60 ~ delPTime, data=d,
    main="File delivery waits",
    xlab="Time of day of delivery", xlim=prettyEdges,
    ylab="Wait time (hours)",
    scales=list(x=list(tick.number=6, cex=0.6))
  ),
  "delWaits.png", h=4, w=6
)
#with graphics delWaits.png timeout 60
```

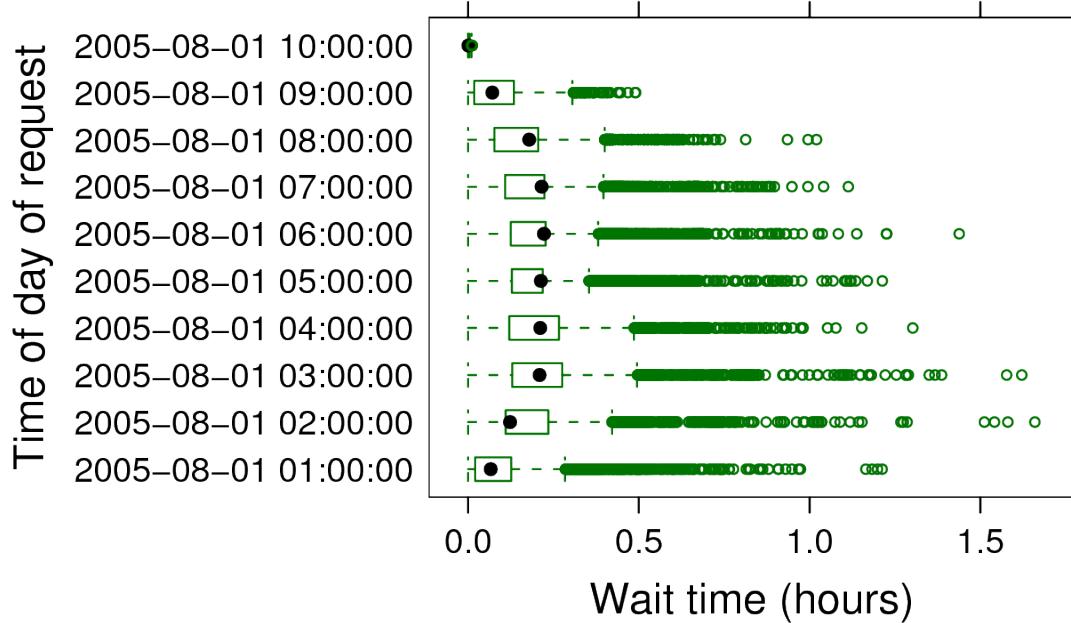
File delivery waits



These plots make the outliers really stand out and hard to see the mean wait times. Let's try to plot those...

```
<R47> mp(
  bwplot( getsPerHourCuts ~ getDur/60/60, data=d,
         main="File delivery waits",
         xlab="Wait time (hours)", ylab="Time of day of request"
       ),
  "waitsPerHourBw.png", h=4, w=6
)
#with graphics waitsPerHourBw.png timeout 60
```

File delivery waits



Again, the outliers dominate the plot. Let's just plot medians and how many are over an hour...

```
<R48> waitMedians = tapply(d$getDur/60, getsPerHourCuts, median)
```

```
<R49> waitMeans = tapply(d$getDur/60, getsPerHourCuts, mean)
```

```
<R50> nOverHour = tapply(d$getDur/60/60 >= 1, getsPerHourCuts, sum)
```

```
<R51> percOverHour = nOverHour / getsPerHour * 100
```

```
<R52> nOverHalfHour = tapply(d$getDur/60/60 >= 0.5, getsPerHourCuts, sum)
```

```
<R53> percOverHalfHour = nOverHalfHour / getsPerHour * 100
```

```
<R54> nUnderMinute = tapply(d$getDur/60 < 1, getsPerHourCuts, sum)
```

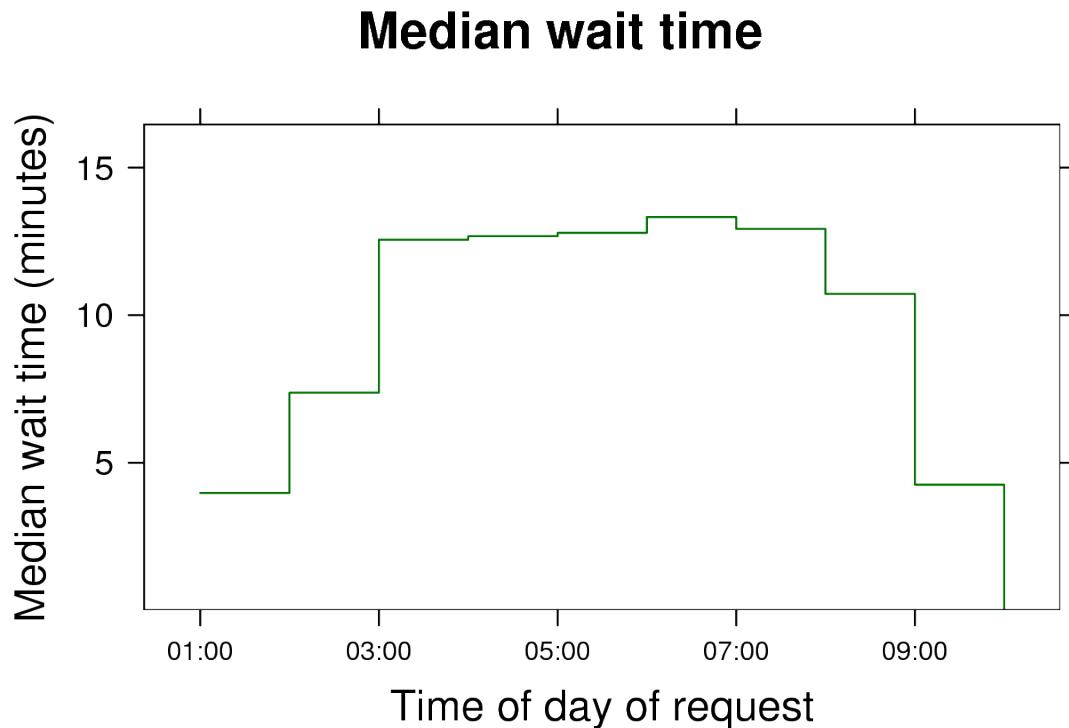
```
<R55> percUnderMinute = nUnderMinute / getsPerHour * 100
```

Let's plot these things

```

<R92> mp(
  xyplot( waitMedians ~ as.POSIXct(names(waitMedians)), type="s",
         main="Median wait time", xlab="Time of day of request",
         ylab="Median wait time (minutes)",
         ylim=c(0, max(waitMeans)*1.2),
         scales=list(x=list(tick.number=6, cex=0.6))
  ),
  "medians.png", h=4, w=6)
#with graphics medians.png timeout 60

```

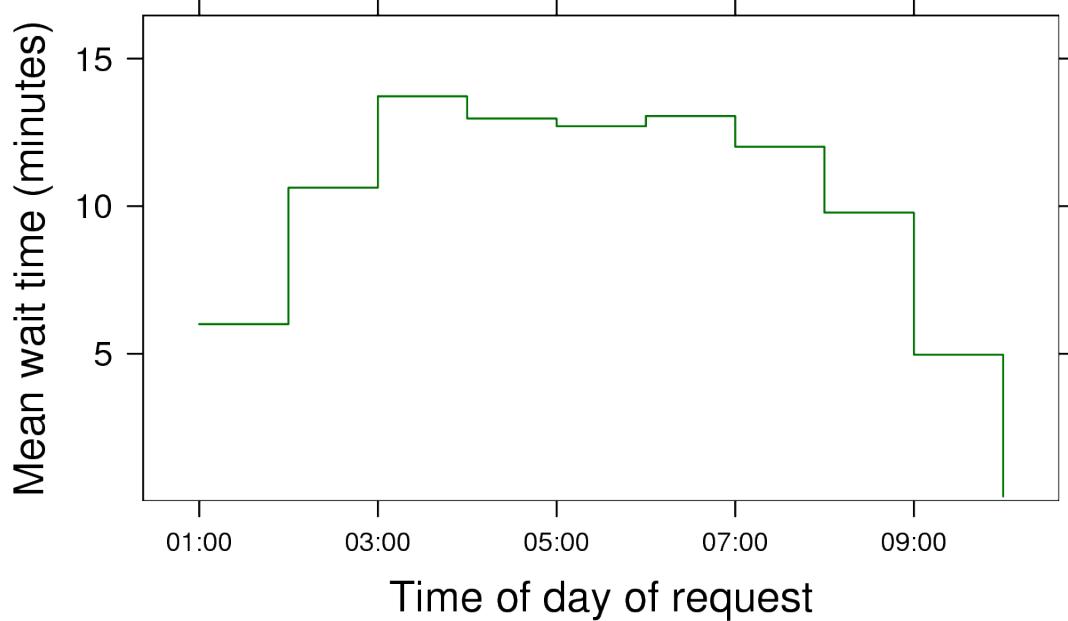


```

<R91> mp(
  xyplot( waitMeans ~ as.POSIXct(names(waitMeans)), type="s",
         main="Mean wait time", xlab="Time of day of request",
         ylab="Mean wait time (minutes)",
         ylim=c(0, max(waitMeans)*1.2),
         scales=list(x=list(tick.number=6, cex=0.6))
  ),
  "means.png", h=4, w=6)
#with graphics means.png timeout 60

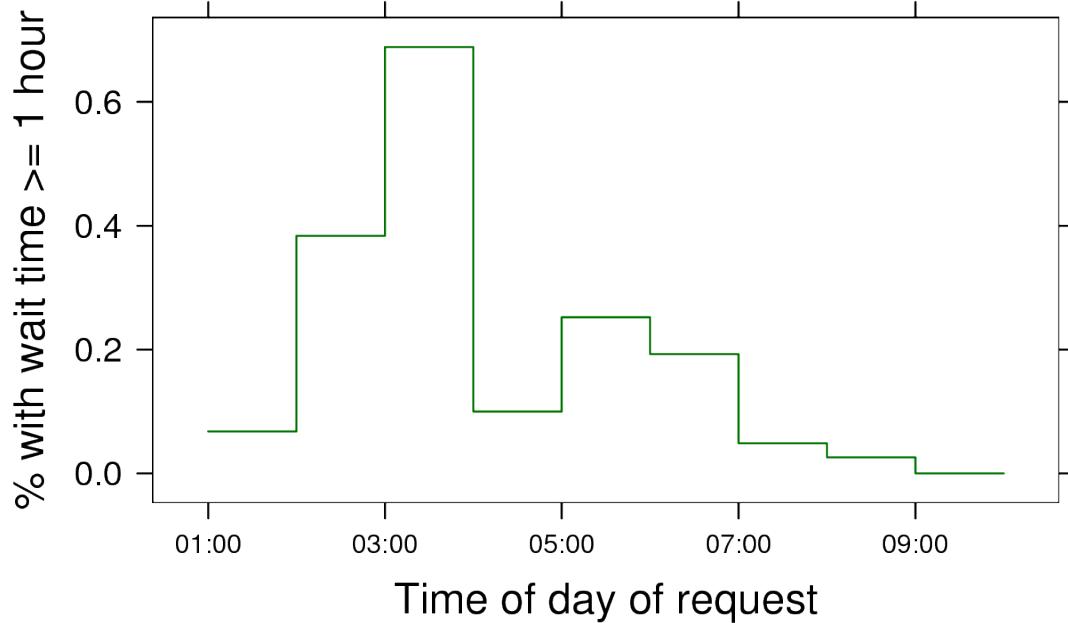
```

Mean wait time



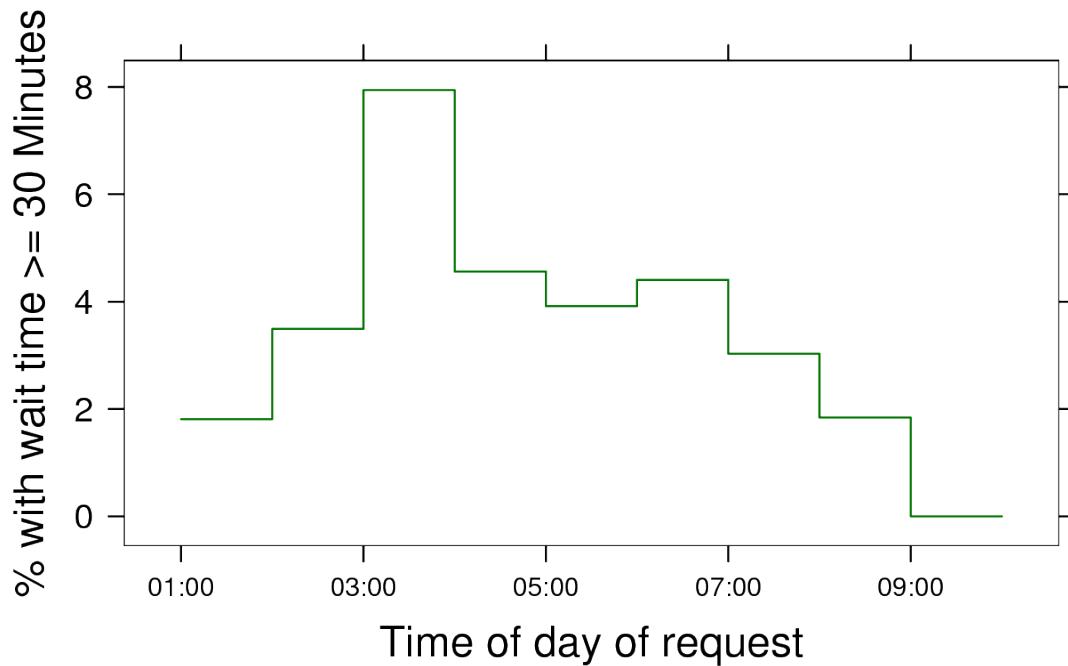
```
<R58> mp(
  xyplot( percOverHour ~ as.POSIXct(names(percOverHour)), type="s",
    main="Percent long wait times", xlab="Time of day of
    request",
    ylab="% with wait time >= 1 hour",
    scales=list(x=list(tick.number=6, cex=0.6))
  ),
  "overHour.png", h=4, w=6)
#with graphics overHour.png timeout 60
```

Percent long wait times



```
<R59> mp(
  xyplot( percOverHalfHour ~ as.POSIXct(names(percOverHalfHour)),
  type="s",
    main="Percent long wait times", xlab="Time of day of
request",
    ylab="% with wait time >= 30 Minutes",
    scales=list(x=list(tick.number=6, cex=0.6))
),
"overHalfHour.png", h=4, w=6)
#with graphics overHalfHour.png timeout 60
```

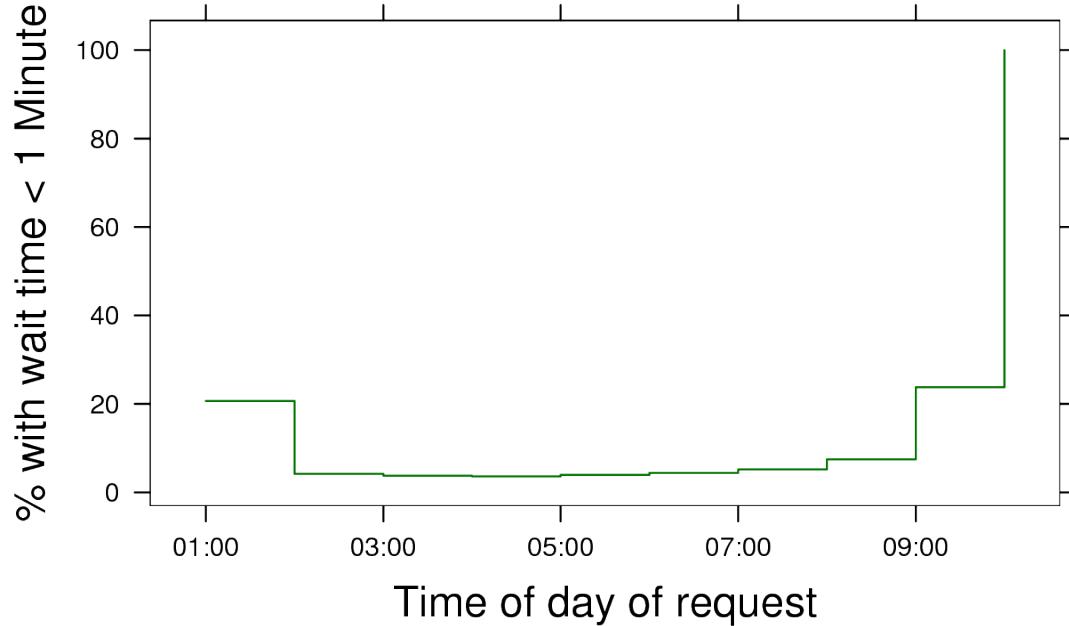
Percent long wait times



Look for short wait times.

```
<R60> mp(
  xyplot( percUnderMinute ~ as.POSIXct(names(percUnderMinute)),
  type="s",
    main="Percent short wait times", xlab="Time of day of
request",
    ylab="% with wait time < 1 Minute",
    scales=list(tick.number=6, cex=0.6)
  ),
  "underMinute.png", h=4, w=6)
#with graphics underMinute.png timeout 60
```

Percent short wait times



A.2.5 Look at number of open requests

Let's look at the number of instantaneous requests that are open at any given time.

Get next file requests...

```
<R61> gnf = data.frame( time=d$getPTime, adj=1 )
```

File deliveries (request fulfilled)

```
<R62> gnc = data.frame( time=d$delPTime, adj=-1 )
```

Join them,

```
<R63> gn = rbind(gnf, gnc)
```

Sort by time,

```
<R64> gn = gn[ order(gn$time), ]
```

Do a running count of # of unfulfilled get next file requests

```
<R65> gn$count = cumsum(gn$adj)
```

```
<R66> gn[1:10,]
```

```

                time adj count
81      2005-08-01 01:05:50   1    1
1       2005-08-01 01:05:51   1    2
81100  2005-08-01 01:05:51  -1    1
110000 2005-08-01 01:05:54  -1    0
161     2005-08-01 01:06:04   1    1
82      2005-08-01 01:06:06   1    2
201     2005-08-01 01:06:07   1    3
82100  2005-08-01 01:06:07  -1    2
161100 2005-08-01 01:06:07  -1    1
2       2005-08-01 01:06:09   1    2

<R74> summary(gn$count)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max.
0.0    707.0  900.0  763.8  906.0  921.0

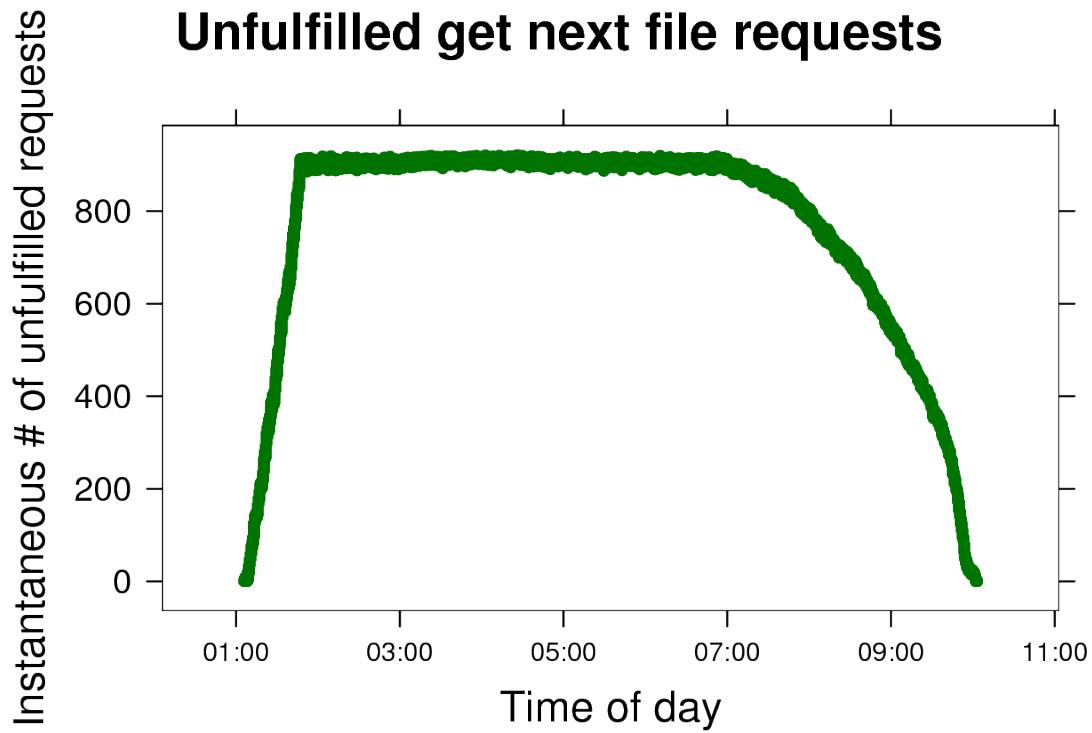
```

Plot it...

```

<R76> mp(
  xyplot( gn$count ~ gn$time, type="p",
          main="Unfulfilled get next file requests",
          xlab="Time of day",
          ylab="Instantaneous # of unfulfilled requests",
          scales=list(x=list(tick.number=6, cex=0.6)),
          xlim=prettyEdges, pex=0.1 ),
  "unfulfilled.png", h=4, w=6 )
#with graphics unfulfilled.png timeout 240

```



I think this just shows the affect of the request wait time being longer than the request rate.